Exome Sequencing Deciphers Rare Diseases

Two years ago, NIH's Undiagnosed Diseases Program began delivering genomics to the clinic on an unprecedented scale. Now, with 128 exomes sequenced and 39 rare diseases diagnosed, the program's success is paving the way for widespread personal genomics while pioneering new techniques for reigning in the "tsunami" of genomics data.

In 2009, a healthy Colombian couple watched their two sons suffer from a mysterious neurological illness, featuring seizures, tremors, and other complications. When the younger son succumbed to the illness at age 13, the family decided to try a new approach. They enrolled their elder son, who will be referred to here as Carlos, in NIH's Undiagnosed Diseases Program, a trans-institute initiative dedicated to deciphering the cause of rare, mysterious health conditions.

The Undiagnosed Diseases Program at NIH started in May 2008 as a pilot program with initial funding of $280,000. Since then, it has grown to include ~75 physicians and scientists from almost every institution at NIH, including endocrinologists, immunologists, oncologists, and cardiologists. The team now works together on ~300 cases with $3.5 million in funding each year.

What sets the program apart from other large clinical projects is their state-of-the-art genetic analyses. The team's Illumina platforms can sequence the entire exome—that is, the 180,000 exons in the human genome—of a patient and his family in only ~9 weeks. In addition, high-resolution microarrays can genotype a million single-nucleotide polymorphisms (SNPs) for each family member, providing information about the remaining 99% of the genome—the introns and non-coding regions. In this way, the Undiagnosed Diseases Program is leading the charge in bringing genomics to the clinic.

So why is this type of clinical genomics not routinely available outside the NIH? Although the price of sequencing has dropped to roughly $15,000 per genome or about $3000 for an exome, data storage and the ensuing analysis remains too costly, too laborious, and frankly too inefficient to put into common practice in the clinic. It's cheap and fast to sequence, but to glean diagnostic information from the data is still quite laborious and costly.

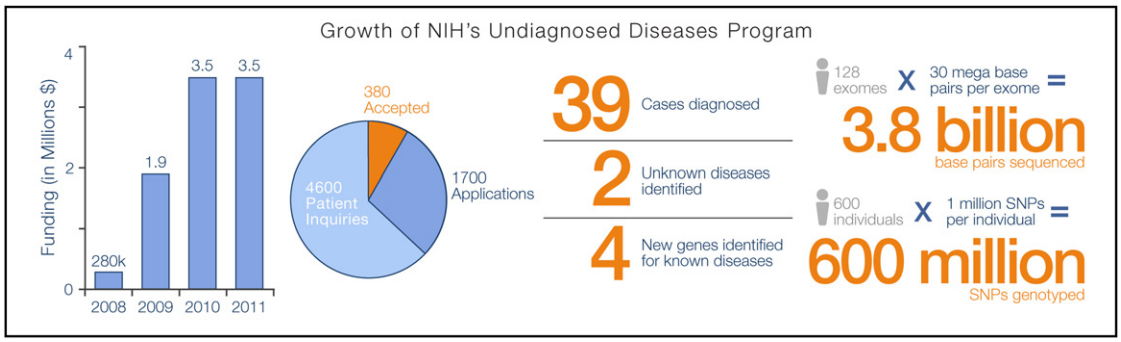
But recent successes by the Undiagnosed Diseases Program are shifting this trend. First, the program is learning how to identify patients who might benefit most from genomic analysis. Second, the team has developed tools for quickly narrowing down the flood of genomics data to a handful of candidate genes for functional tests. With these new strategies, the program has made 39 diagnoses, including the team's first reported discovery of a new disease, which was published last month in *The New England Journal of Medicine.*

Through the program, "scientists and clinicians will get an understanding of the difficulties involved in next-gen sequencing, as well as the benefits; of what cases are appropriate for genomic study and of what filters should be applied to the data, " says William Gahl, the program's director and clinical director of the National Human Genome Research Institute.

Family Matters First

Before Carlos and his family arrived at the NIH Clinical Center, their first major hurdle was acceptance into the program. The NIH has received ~4600 inquiries for the program and ~1700 applications, but it has enrolled only 380 cases. In general, the program selects patients whose maladies appear genetic in origin, based on their family history or a sign that their symptoms have a shared underlying cause.

Typically, the NIH team first attempts to diagnose illnesses by looking for known genetic markers and using standard molecular and biochemical tests that are commercially available. However, if they have no good leads on the disease and therefore no candidate genes to search for, they'll look to a patient's whole exome. An entire exome yields an overload of information, however, and must be carefully edited. Exome sequences typically vary at about



20,000 places, explains Thomas Markello, a clinician involved with the program at NHGRI. This vast variation makes it nearly impossible to find 1–2 rare alleles underlying a mysterious disease. In other words, the nucleotide haystack is just too big.

After some trial-and-error, the team has developed a strategy to quickly shrink the haystack: they eliminate many variants by comparing the patient's DNA with genomic information from the family. This added information enhances the signal-to-noise ratio and allows them to reduce the number of candidate genes. "Data reduction is pretty much everything," Markello says.

"It's simply not worth doing whole exome sequencing on a single individual," says Gahl, "when the clinical manifestations don't point to a specific group of genes." With family information, the team can apply classical pedigree tools reminiscent of population genetics in the 1970s to filter out mutations that don't follow predicted modes of Mendelian inheritance. And DNA from healthy family members allows them to eliminate harmless variations that run in the family but don't lead to the disease.

"Family data makes a fundamental difference," says David Adams, an NHGRI clinician who worked on Carlos's case. "You don't just need family history; we've learned you need family DNA to succeed."

So for Carlos's case, the geneticists sequenced his exome, as well as that of his parents and his deceased brother. In general, the team gets ~88% of the complete exome with 99.9% confidence. The exome data from Carlos and his family initially generated ~120,000 variants for program geneticists to sift through. Remarkably, the team then narrowed this list to three candidate genes using a series of software programs developed internally at the NIH.

The geneticists first align the patient's exome with a reference sequence (typically the one generated in the Human Genome Project) and exclude unlikely candidate mutations according to a kill-list of variants present in more than 1% of exomes and genomes stored in databases from two separate projects, NIH's ClinSeq and the 1000 Genomes Project. Mutations are then ranked by how severely they alter the coding sequence. For example, a mutation encoding a stop codon is considered more severe than one with no effect on the resulting amino acid.

Another in-house software program then eliminates mutations if their inheritance pattern doesn't match that of the disease. For example, Carlos's disease appeared to follow a recessive mode of inheritance, and thus, his team kept only the variants that behaved accordingly. Finally, the team exports the results to their program VarSifter, which lists the candidate variants in order from least to most likely.

Currently, some of this software is available to the scientific community, and according to NHGRI's Nancy Hansen, the entire suite may be released once it's optimized so that clinicians and scientists might follow the lead of the Undiagnosed Diseases Program.

Success Stories Accumulating

Although the program's strategy is still evolving, this "whittling down" approach is producing results. At the end of 2010, the program identified the root of Carlos's neurological symptoms, and a report on the teen's diagnosis is pending publication. And data from the Undiagnosed Diseases Program team also helped Manfred Boehm at the National Heart, Lung and Blood Institute quickly locate the mutation behind a mysterious vascular calcification disorder. The team handed him a list of 100 candidate genes, and within ~1 month he had pinpointed mutations in the *NT5E* gene as the causal culprit.

NT5E encodes a membrane-bound nucleotidase involved in extracellular ATP metabolism. Boehm's study identified in three families numerous mutations, which destroy the activity of the nucleotidase. By targeting this enzyme, clinicians may soon be able to treat a disease that didn't even have a name a few months ago. The study was published last month in *The New England Journal of Medicine*.

In contrast, Boehm says, a couple of years ago, NIH geneticists handed him a list of 2000 genes potentially underlying a known disorder, and a year passed before he could identify the causal mutation.

Thus far, the program has diagnosed 39 cases—3 of which are neurological and muscular diseases discovered through exome analysis. Another 3 were discovered with SNP analysis, whereas the remaining cases were diagnosed with commercial tests, including the identification of the rare disease congenital disorder of glycosylation type 2B.

The Undiagnosed Diseases team also counts concise lists of candidate mutations as partial successes. These "short lists" are handed off to bench scientists for functional studies on how the mutations contribute to disease. Right now, these collaborations are currently all within the NIH. But program leaders say they'll work with basic researchers outside of the NIH once they've established a portal for collaboration.

Marjan Huizing, a metabolic disorders researcher at NHGRI, is not involved with the program but says she's "piggybacking" off their exomic techniques. She studies endocytic trafficking defects that lead to albinism, bleeding, and infections.

Traditional candidate gene approaches have frustrated her team. They spent 12 years screening patients for candidate genes and never found the right one. So she decided to follow in the program's lead and analyze the exomes of three patients. Right off the bat, she identified the causal mutation underlying two of the cases. For the third, more difficult case, she's planning to sequence the patient's parents.

In the meantime, geneticists are keeping an eye on the program for clues about the nature of rare diseases and whether sequencing alone can identify "medically actionable alleles," says Harvard genomicist George Church. "If that turns out to be true, it's a paradigm shift."

Although the clinicians at the Undiagnosed Diseases Program echo Church's curiosity, they keep the focus tightly on their patients. The thrill of a diagnosis through genomics is a triumph for scientists and clinicians, but it means even more to patients and their loved ones, who have sought explanations for years.

Still, it is a bittersweet success. Carlos now has a name to call his malady, but there remains no cure or powerful drug

to combat it. ''Even when we get lucky, the best we can do is offer hope to the patient that down the road, someone might use the diagnosis to develop a treatment,'' says John Gallin, director of the NIH Clinical Center. ''Our patients have referred to this as the House of Hope, but from the perspective of a care provider, I wish I could do even more.''

Amy Maxmen
New York, NY, USA
DOI 10.1016/j.cell.2011.02.033